ChinaXiv合作期刊

主题标引文献的语义关系发现研究*

李晓瑛 夏光辉 李丹亚
(中国医学科学院医学信息研究所 北京 100020)

摘要: 【目的】利用文献的主题标引结果, 发现其中隐含的重要语义关系。【方法】基于 MEDLINE 数据库中的生物医学主题标引文献, 提出一种语义关系发现算法, 涉及主题词组配原则、主题标引规则以及基于加权标引词和关系出现频次的优化方法等多个环节。【结果】收集疾病与症状方面的实验数据对算法进行实验验证, 并结合领域专家审核, 结果表明本文所发现语义关系的准确率可达到 95%以上。【局限】本文所研究的语义关系发现算法仅适用于具有主题标引结果的文献。【结论】从大规模生物医学主题标引文献中发现中英文两种语言的语义关系是有效可行的, 对其他领域语义关系的发现具有极高的借鉴意义。

关键词: 语义关系发现 标引文献 组配原则 阈值
分类号: G250

chinaXiv:201711.02059v1

1 引 言

随着科学研究与数字出版的快速发展与推进, 公开出版的文献已然成为重要的数据来源。据统计, 截至 2015 年 5 月, 美国 MEDLINE 数据库所收录的生物医学期刊文献总量已超过 2 200 万篇, 并且每年以 70 万篇的速度增长[1]。显然, 对科研人员而言, 欲从如此庞大的文献数据中及时获取新思想、新方法、新成果, 无疑是一件极具挑战性的工作。

为了有效地从科学研究中发现新知识并付诸于实践, 越来越多的学者投入到结构化或半结构化文本信息的自动抽取研究中, 即信息抽取(Information Extraction, IE)。一般而言, IE 过程涉及两项重要任务: 命名实体识别(Named Entity Recognition, NER)及关系抽取(Relation Extraction, RE)[2]。前者旨在从文本中识别实体类型(如人名、地名、组织机构名、疾病、药物、基因、蛋白等)及相应的实体名称, 而后者则侧重确定两个实体(或概念)间重要的语义关系。例如, 在处理"利心平(Nifedipine)是治疗高血压(Hypertension)的常见药物"、"肺炎(Pneumonia)的症状包括呼吸困难(Dyspnea)等"时, 关系抽取 RE 的任务就是发现"利心平"与"高血压"之间的"治疗"关系、"肺炎"与"呼吸困难"之间的"症状"关系。而诸如此类的关系抽取成果, 具有较高的实际应用价值。自动问答等信息检索系统即为一个典型应用, 用于回答形如"哪种药物可用于治疗高血压"、"肺炎都有哪些症状"的提问。而在叙词表、本体等知识组织系统、领域知识库及语义网的构建中, 关系抽取能够丰富概念间的语义关系, 增加关系实例, 扩充知识结构。

2 相关研究

关于关系抽取的研究, 至今已取得一定的研究成果。而依据其发现一对实体间语义关系的基本原理, 这些研究大体可分为三类: 基于模式匹配、基于机器学习及基于词表的方法[3]。

(1) 基于模式匹配的 RE 研究首先利用语言学知识或领域知识生成若干关系模版, 之后再将待处理句子与模版逐一进行匹配。一旦匹配成功, 则认为该句具有模版特征, 从而认定句中实体间的语义关系[4-6]。例如, 从"A 即/亦即/或/或称/也称 B"模版中, 发现 A

通讯作者: 李晓瑛, ORCID: 0000-0003-4407-6616, E-mail: lixiaoying@imicams.ac.cn。

*本文系国家社会科学基金项目"基于复杂网络的公众健康知识网络构建研究"(项目编号:15CTQ020)和中央级公益性科研院所基本科研业务费项目"生物医学术语服务系统建设关键问题研究"(项目编号: 15R0109)的研究成果之一。

chinaXiv:201711.02059v1

与B之间的同义关系。其中，关系模版通常由领域专家手工生成，或由计算机程序依据一定规则从语料中自动产生。

(2) 基于机器学习的算法将自然语言处理技术应用于关系抽取任务中，包括有监督、无监督及弱监督三种方式。有监督的机器学习中，领域专家事先标注出语料库中的语义关系，RE算法则依据词法、语法及语义特征训练分类器，并将其用于发现待处理文本中的语义关系[7-9]。然而生成语料库往往需要领域专家的参与，不仅费时费力，而且较难扩展到其他领域中，因此无监督的关系抽取研究应运而生。该方法首先自动抽取句中实体间的语义关系，而后再对大量的关系进行聚类[10-11]；相对而言，在处理专业领域文献时，其聚类结果还有待进一步优化完善。弱监督的关系抽取技术综合考虑了上述两种方法的优缺点，主要改进之处在于利用领域语料库中句子的特征自动训练分类器，在提高RE结果专业适应性的同时，减少领域专家的人工参与[12-13]。总体而言，基于机器学习的关系抽取方法因其具有较高的计算效率和较少的领域知识和专家参与，目前较多地用于大规模通用文本处理中。

(3) 基于词表的语义关系抽取方法，从已有的语义词典或成熟的领域本体中获取实体之间的语义关系[14-15]，例如从WordNet中发现两个词之间的上下位关系；鉴于词典和本体中的内容结构均已通过编制者的审定，该方法所抽取的语义关系一般具有较高的准确性。然而，其局限性也相当明显，因为词表收词量十分有限，且很难实时更新。

关系抽取也称关系发现，一般而言二者无严格区别；但在信息领域中，前者多指从一个句子或其相邻上下文中确定(Identify)实体间的关系，而后者的范围可扩展至整篇文献或文本。与上述各种语义关系抽取方法不同，本研究从多年生物医学领域主题词表编制及文献标引经验出发，探讨基于主题标引文献的语义关系发现算法，旨在从大规模的文献数据中发现重要的语义关系，为构建基于语义关系的信息检索系统、知识组织系统、领域知识库及语义网提供数据基础。相对而言，这种基于主题标引文献的语义关系发现算法具有较高的准确性，因为文献标引主题词一般由标引员给出，或由程序自动计算后人工审核，在一定程度上降低了错误率；此外，本研究提出一种基于加权标引词和关系出现频次的优化方法，进一步提高算法的准确率。值得提出的是，虽然本研究针对生物医学主题标引文献展开，但这种基于主题词组配原理及主题标引规则的语义关系发现机制，对其他领域而言，具有极高的借鉴意义。

3 算法发展基础

3.1 《医学主题词表》

《医学主题词表》(Medical Subject Headings, MeSH)由美国国立医学图书馆(National Library of Medicine, NLM)负责编制及更新维护[16]，是目前公认的最权威的生物医学主题词表，广泛用于生物医学文献的标引与检索、图书编目等基于生物医学主题词描述文献实质内容的数据库中。2016版MeSH，共包含疾病、病因、体征、药物等在内的27 883个主题词，以及畸形、化学诱导、病因学、并发症、药物疗法等82个副主题词，主题词与副主题词组配使用，副主题词对主题词起到限定或复分的作用，使主题词具有更高的专指性。例如“肾发育不全”，在输入主题词“肾”后，选择副主题词“畸形”表示发育不全。另外，MeSH亦是一部规范化的可动态扩展的生物医学领域叙词表；中国医学科学院医学信息研究所在对MeSH进行汉化实现中英文双语对照的基础上，增加了《中国中医药学主题词表》相关内容，形成《中文医学主题词表》(Chinese Medical Subject Headings, CMeSH)[17]，全面用于中文生物医学文献的标引、编目与检索。对于本研究而言，MeSH与CMeSH提供了中英文两种语言的生物医学实体(概念)名称，从而为所发现的语义关系进行中英文转换奠定了基础。

3.2 MEDLINE生物医学主题标引文献

MEDLINE是由NLM开发的大型开放性生物医学文献数据库[1]，使用MeSH词表对生物医学文献进行主题标引与检索，公众可自由获取全文文献及其基于MeSH词表的主题标引结果(对应的检索系统为PubMed)。如图1所示，一篇论证呼吸困难为肺炎症状的文章(PMID为文章编号)，依据标引规则并利用MeSH词表进行主题标引后，标引词(MH)包含“肺炎/并发症”、“呼吸困难/病因学”。

ChinaXiv合作期刊

chinaXiv:201711.02059v1

```
PMID- 20735868
OWN - NLM
STAT- MEDLINE
DA  - 20100825
DCOM- 20100920
TI  - [A man with exercise related shortness of breath].
PG  - A1102
AB  - A 51-year old male was admitted to the hospital with complaints of fever, a
      productive cough and exercise-related shortness of breath. These complaints were
      caused by a pneumatocele, which was successfully treated with antibiotics.
MH  - Anti-Bacterial Agents/therapeutic use
MH  - Cough/diagnosis/etiology
MH  - Dyspnea/*diagnosis/*etiology
MH  - *Exercise/physiology
MH  - Humans
MH  - Male
MH  - Middle Aged
MH  - Pneumonia/*complications/*diagnosis/drug therapy
```

图 1　MEDLINE 主题标引文献示例

3.3　基于 MEDLINE 生物医学主题标引文献的语义关系发现规则

基于 MeSH 词表的生物医学文献主题标引的组配原则以及 MEDLINE 数据库所提供的主题标引文献，为本研究发现生物医学实体(或概念)间的语义关系(特别是与公众健康密切相关的疾病知识)提供了一定的理论与数据基础。例如，从图 1 所示的文献标引主题词中，可发现“肺炎”与“呼吸困难”之间的“临床发现(即症状)”关系。具体而言，这些语义关系不仅包括疾病与体征之间的临床发现关系(即症状)，还有疾病与化学物质、基因、微生物之间的引发关系(即病因)，疾病与药物之间的治疗关系，疾病与诊断技术和方法之间的诊断关系，疾病间的并发关系，肿瘤间的继发关系(继发关系一般仅针对肿瘤)等，相应的主题词与副主题词的组配原则如表 1 所示。例如，通过在一篇文献中同时出现的一组主题标引结果“疾病/并发症”、“症状与体征/病因学”，则揭示了疾病与症状、体征之间的临床发现关系。

表 1　基于 MEDLINE 主题标引文献的语义关系发现规则

语义关系	主题标引结果 1		主题标引结果 2	
	主题词	副主题词	主题词	副主题词
临床发现(症状)	疾病	并发症	症状与体征	病因学
	疾病	病因学(或化学诱导)	化学物质	副作用(或中毒)
引发(病因)	疾病	遗传学	基因	
	疾病	微生物	微生物	
(药物)治疗	疾病	药物疗法	药物	治疗应用(或投药&剂量)
诊断	疾病	诊断	诊断技术和方法	方法
并发	疾病 1	并发症	疾病 2	并发症
继发	肿瘤 1	病理学	肿瘤 2	继发性

3.4　语义关系发现优化研究

在对文献进行主题标引时，标引员通常采用为最能表达文献主题内容的标引词打星号(或 IM)的方式区分标引词的权重，即加权标引[18]；带有星号的标引词为文献重点讨论内容，其重要程度也最高；进而，基于带星号的标引词所推导出的语义关系不仅关键而且准确，因为在经过人工标引或自动标引及人工审核后，文献最核心主题标引词一般很少标错。此外，为了杜绝个别作者编撰数据、撰写不真实的学术文章，本研究进一步引入发现某一对具体关系的文献数(即关系的出现频次)对所发现的语义关系进行优化，并依据统计学原理设定相应的阈值作为控制参数，以提高发现结果的可靠性与准确性。

3.5　基于 MEDLINE 生物医学主题标引文献的语义关系发现算法

根据上述 MeSH 词表中主题词与副主题词组配原则、生物医学文献主题标引规则、MEDLINE 所提供的主题标引文献数据格式以及语义关系发现优化机制，本文提出一种基于生物医学主题标引文献的语义关系发现算法，其基本思想如图 2 所示。

(1) 从 MEDLINE 数据库中获取生物医学主题标引文献，并记录每篇文章编号 PMID 及所有的主题标引词 MH;

(2) 对主题标引词 MH 进行筛选，仅保留带星号的加权标引词;

(3) 逐一将每篇文章与语义关系发现规则(见表 1)进行匹配，保留符合主题词与副主题词组配原则的文献及主题标引词 MH，对其余文献进行滤除;

(4) 依据语义关系发现规则，提取语义关系三元组(概念 1、语义关系类型、概念 2)，并记录相应的文章编号 PMID;

(5) 按照语义关系三元组进行聚类，统计相应的文章个数，作为该关系的出现频次;

(6) 根据一定的统计学原理，选择有意义的阈值，

ChinaXiv合作期刊

chinaXiv:201711.02059v1

对所发现的语义关系进行优化；关系出现频次小于阈值的语义关系三元组，将被滤除；

(7) 经过优化后的语义关系三元组，将被作为最终结果输出。

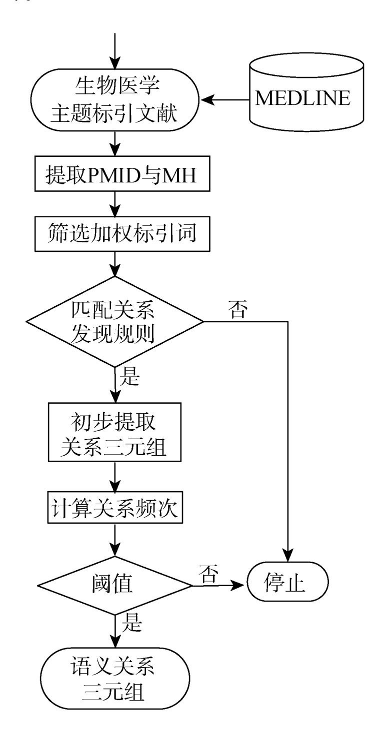


图 2　基于 MEDLINE 生物医学主题标引文献的语义关系发现流程

4　实验与讨论

本文以呼吸道疾病与症状、体征之间的临床发现关系为例，系统地阐述基于主题标引文献的语义关系发现全过程，包括从 MEDLINE 获取数据、算法优化中阈值的选取等多个环节；并邀请领域专家对实验所发现的语义关系逐一进行审核，在验证算法准确率的同时，深入分析实验结果。

4.1　数据获取与算法实现

鉴于本实验以呼吸道疾病与症状、体征之间的临床发现关系为例测试语义关系发现算法，因此仅需获取 MEDLINE 数据库中论述呼吸道疾病相关症状与体征的文献集合，并非全部文献。而 PubMed 检索平台可根据 MeSH 词表中主题词与副主题词组配原则设置检索条件，并支持二次检索，为本研究从 MEDLINE 数据库获取符合语义关系发现规则的生物医学主题标引文献集合给予了保障。公开获取相应文献数据集的基本步骤如下：

(1) 检索含指定 MeSH 主题词的文献集合。首先设置过滤条件为 MeSH 数据，并输入 MeSH 主题词“呼吸道疾病(Respiratory Tract Diseases)”，表明选择以 MeSH 词表进行主题标引、主题词含“Respiratory Tract Diseases”及其下位的文献，如图 3 所示：

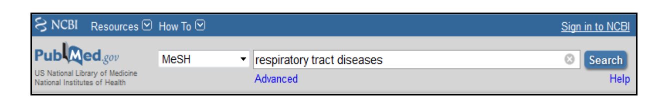


图 3　在 PubMed 中检索含 MeSH 主题词的文献集合示例

(2) 限定主题词所组配的副主题词；在返回页面中，选择主题词“Respiratory Tract Diseases”，选取副主题词“并发症(complications)”，如图 4 所示，并选择附加条件“限制到 MeSH 主要主题(Restrict to MeSH Major Topic)”。

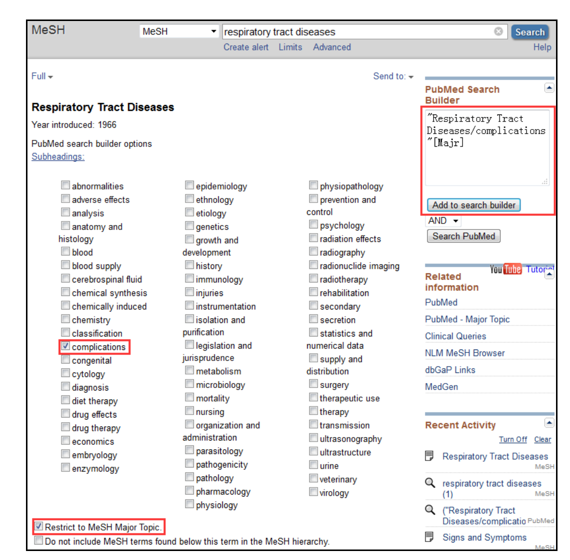


图 4　在 PubMed 中限定主题词所组配的副主题词示例

(3) 以二次检索方式，选择同时含有另一主题词的文献集合。在所返回的页面中，重复上述两个步骤，即设置主题词“症状与体征(Signs and Symptoms)”，限定相应的副主题词为“病因学(etiology)”，并以两次检索条件进行检索；

(4) 获取含主题标引词的文献集合。在返回的结果页面中，选择批量将文献集合下载到本地(File)，并指定格式为含主题标引词(MEDLINE)，点击下载(Create File)后，可将相应的文献数据集保存到指定的

ChinaXiv合作期刊

chinaXiv:201711.02059v1

本地路径中。至此，完成数据获取，如图 5 所示。

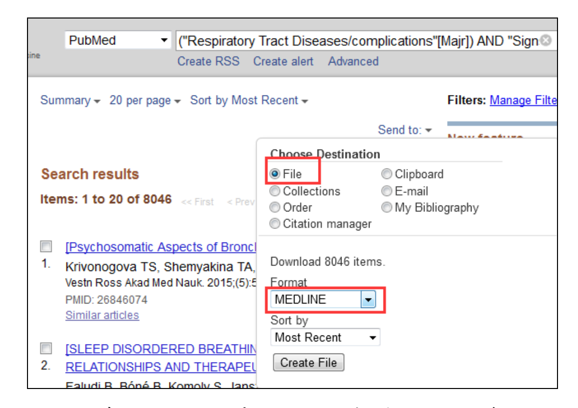


图 5　在 PubMed 中批量下载符合指定条件的生物医学主题标引文献数据集示例

不难看出，上述数据获取过程更多地涉及查找文献时的主题词与副主题词组配环节，而最终在 PubMed 系统使用的检索表达式为：

("Respiratory Tract Diseases/complications" [Majr]) AND "Signs and Symptoms/etiology" [Majr]

考虑到 MEDLINE 数据库每月更新，相比人工多次手动获取数据，基于检索表达式的自动处理算法更受青睐。在获取 MEDLINE 主题标引文献后，本研究逐步实现了语义关系发现算法中的提取 PMID 与 MH、筛选加权标引词、匹配关系发现规则、初步提取关系三元组及计算关系频次共 5 个重要环节。本次实验具体的数据结果为，从 MEDLINE 数据库获取 8 046 篇文献，从中初步提取出 6 468 对关系三元组。

4.2　阈值的选取

进一步分析关系三元组及语义关系出现频次，将关系出现频次作为横坐标，符合该频次的关系三元组个数作为纵坐标，即得到如图 6 所示的语义关系出现频次分布。其中，语义关系出现频次最小为 1，最大为 102；相应地，频次为 1 的关系三元组共有 4 987 对，频次为 102 的关系三元组只有 1 对。另外，图 5 表明随着语义关系出现频次的增加，相应的关系三元组数目呈现递减，其趋势逼近线性分布。

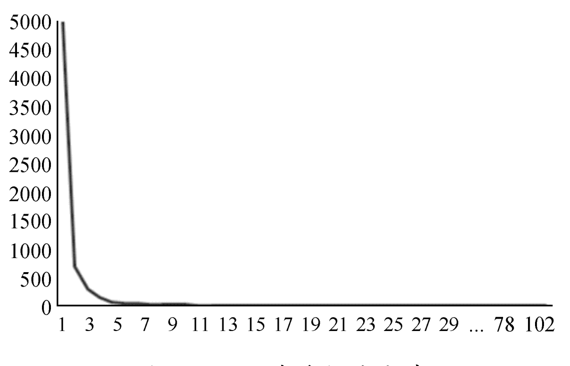


图 6　语义关系频次分布

在统计学中，一般有意义的线性函数统计指标包括均值、中位数、众数等。其中均值也称平均数，指一组数据中所有数据之和与数据个数之比值，是一项反映数据集中趋势的重要指标，用于体现这组数据的一般情况和平均水平。在本研究中，选取代表语义关系平均出现频次的均值作为优化语义关系发现结果的阈值，而出现频次大于阈值的关系三元组，因具有较高的出现频次，将作为语义关系最终发现结果。阈值 Th 的计算公式如下：

$$Th = \frac{\sum_{i=1}^{N} f(i)}{M} \quad (1)$$

其中，N 指语义关系出现频次，f(i)指经统计后出现频次为 i 的关系三元组个数，M 为按主题词去重之后带有出现频次的语义关系三元组总数。根据公式(1)所计算的阈值为 1.814，而基于此阈值进行优化后的语义关系共有 1 481 条。

4.3　专家审核

为了验证所提出的语义关系发现算法的准确性，两位领域专家对实验中算法自动发现的 1 481 条语义关系逐一进行审核，并对其中的 33 条关系不予认可，表明本次实验中语义关系发现算法的准确率为 97.8%。而后对专家不予认可的语义关系进行详尽分析，发现主要原因为标引时所用的主题词过于宽泛，如“癌(Carcinoma)”与“咯血(Hemoptysis)”，这种情况出现的比例高达 75.8%。

4.4　结果讨论

经算法优化与专家审核后，本实验最终共发现 1 448 条关于呼吸道疾病与症状、体征之间的语义关系。在 MEDLINE 数据库中，这些语义关系的出现频次均不低于 2 次，出现频次最高(172 次)的语义关系为“哮喘(Asthma)”与“咳嗽(Cough)”。类似地，利用基于 MEDLINE 生物医学主题标引文献，可发现疾病的病因、治疗、并发症、继发症等语义关系。另外，MeSH 与 CMeSH 中英文对照的词表数据能够支持直接将从 MEDLINE 文献中所发现的英文语义关系转换为中文

ChinaXiv合作期刊

chinaXiv:201711.02059v1

格式，有助于开展基于中英文语义关系的应用实践。同时，相比已有的基于模式匹配、基于机器学习及基于词表的方法，本研究所提出的基于主题标引文献的语义关系发现方法，不仅省去了较多的领域专家干预、语料库选取及算法训练等环节，而且可用于从大规模的生物医学主题标引文献中发现具有较高准确性的中英文两种语言的语义关系。

5 结　语

大规模准确可靠的语义关系对自动问答等信息检索系统、知识组织系统、领域知识库及语义网的构建具有至关重要的影响。本研究立足主题词组配原理及主题标引规则，提出一种基于生物医学主题标引文献的语义关系发现算法，并从加权标引词和关系出现频次等多角度对算法进行优化。通过从 MEDLINE 获取实验数据进行验证，并经领域专家审核，获得满意的准确率，其结果可投入应用实践。最后，尽管本文选取生物医学领域发展算法并收集数据进行验证，但这种基于主题词组配原理及主题标引规则从主题标引文献发现语义关系的算法原理，对其他领域开展基于主题标引文献的语义关系发现研究具有极高的借鉴意义。

参考文献：

[1] U.S. National Library of Medicine. MEDLINE Fact Sheet [EB/OL]. [2016-03-01]. http://www.nlm.nih.gov/pubs/factsheets/medline.html.

[2] 黄勋，游宏梁，于洋. 关系抽取技术研究综述[J]. 现代图书情报技术，2013(11): 30-39. (Huang Xun, You Hongliang, Yu Yang. A Review of Relation Extraction [J]. New Technology of Library and Information Service, 2013(11): 30-39.)

[3] 徐健，张智雄，吴振新. 实体关系抽取的技术方法综述[J]. 现代图书情报技术，2008(8): 18-23. (Xu Jian, Zhang Zhixiong, Wu Zhenxin. Review on Techniques of Entity Relation Extraction [J]. New Technology of Library and Information Service, 2008(8): 18-23.)

[4] Yu H, Hatzivassiloglou V, Friedman C, et al. Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles [C]. In: Proceedings of the 2002 AMIA Annual Symposium. 2002.

[5] 宋锐，林鸿飞，常富洋. 中文比较句识别及比较关系抽取[J]. 中文信息学报，2009, 23(2): 102-122. (Song Rui, Lin Hongfei, Chang Fuyang. Chinese Comparative Sentences Identification and Comparative Relations Extraction [J]. Journalof Chinese Information Processing, 2009, 23(2): 102-122.)

[6] 韩红旗，徐硕，桂婕，等. 基于词形规则模板的术语层次关系抽取方法[J]. 情报学报，2013, 32(7): 708-715. (Han Hongqi, Xu Shuo, Gui Jie, et al. Term Hierarchical Relation Extraction Method Based on Morphology Rule Template [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(7): 708-715.)

[7] Reichartz F, Korte H, Paass G. Dependency Tree Kernels for Relation Extraction from Natural Language Text [A]. // Machine Learning and Knowledge Discovery in Databases [M]. Springer Berlin Heidelberg, 2009.

[8] 孙霞，董乐红. 基于监督学习的同义关系自动抽取方法[J]. 西北大学学报：自然科学版，2008, 38(1): 35-39. (Sun Xia, Dong Lehong. Automatic Extraction of Synonymy Relation Using Supervised Learning [J]. Journal of Northwest University: Natural Science Edition, 2008, 38(1): 35-39.)

[9] 庞晓东. 基于监督学习的校友实体关系抽取研究[D]. 天津：南开大学，2012. (Pang Xiaodong. Research on the Alumni Entity Relation Extraction Using Supervised Learning[D]. Tianjin: Nankai University, 2012.)

[10] Rozenfeld B, Feldman R. High-Performance Unsupervised Relation Extraction from Large Corpora[C]. In: Proceedings of the 6th International Conference on Data Mining. 2006: 1032-1037.

[11] 马超. 基于 Web 信息使用改进的无监督关系抽取方法构建交通本体[J]. 计算机系统应用，2015, 24(12): 273-276. (Ma Chao. Using Improved Unsupervised Relation Extraction Method to Construct Traffic Ontology Based on Web [J]. Computer Systems & Applications, 2015, 24(12): 273-276.)

[12] Zhang Z. Weakly-Supervised Relation Classification for Information [C]. In: Proceedings of the 13th ACM International Conference on Information & Knowledge Management. 2004.

[13] Fan M, Zhao D, Zhou Q, et al. Distant Supervision for Relation Extraction with Matrix Completion [C]. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Maryland, USA. 2014.

[14] Sabou M, Mathieu A, Motta S. CARLET: Semantic Relation Discovery by Harvesting Online Ontologies [C]. In: Proceedings of the 5th European Semantic Web Conference. 2008.

[15] 李熙，徐德智. 基于 WordNet 的概念语义相似度研究[J]. 湖南科技学院学报，2008, 29(12): 115-116. (Li Xi, Xu

ChinaXiv合作期刊

Dezhi. Concept Semantic Similarity Researching Based on WordNet [J]. Journal of Hunan University of Science and Engineering, 2008, 29(12): 115-116.)

[16] U.S. National Library of Medicine. MeSH Browser [EB/OL]. [2016-03-01]. http://www.nlm.nih.gov/mesh/MBrowser.html.

[17] 中国医学科学院医学信息研究所. 中文医学主题词表[EB/OL]. [2016-03-01]. http://cmesh.imicams.ac.cn/index. (Institute of Medical Information, Chinese Academy of Medical Sciences. Chinese Medical Subject Headings [EB/OL]. [2016-03-01]. http://cmesh.imicams.ac.cn/index.)

[18] 肖晓旦. 生物医学文献主题标引[M]. 长沙: 湖南科学技术出版社, 2005: 65-68. (Xiao Xiaodan. Biomedical Literature Subject Indexing [M]. Changsha: Hunan Science & Technology Press, 2005: 65-68.)

作者贡献声明:

李晓瑛: 论文撰写, 算法实现, 数据验证;
夏光辉: 收集实验数据, 参与算法设计;
李丹亚: 提出研究思路, 设计研究方案。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: lixiaoying@imicams.ac.cn。
[1] 李晓瑛, 夏光辉, 李丹亚. Signs and Symptoms about Respiratory Tract Diseases from Indexed Biomedical Papers.xls. 从生物医学主题标引文献发现的与呼吸道疾病相关的症状数据.

收稿日期: 2016-03-09
收修改稿日期: 2016-04-15

Finding Semantic Relations Among Subject Indexed Papers

Li Xiaoying Xia Guanghui Li Danya
(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

Abstract: [**Objective**] This paper tries to identify important and implicit semantic relations among the subject indexed papers. [**Methods**] Based on the subject indexed biomedical papers from MEDLINE, we proposed an algorithm consisting of subjects coordinating and indexing rules, as well as optimization rules for weighted indexing results and relation occurrences. The new algorithm was then examined with experimental disease data. [**Results**] With the help of domain experts' verification, the precision of the new algorithm was higher than 95%. [**Limitations**] The proposed method was only appropriate for papers with subject indexing. [**Conclusions**] The proposed algorithm can be used to identify semantic relations among English and Chinese subjects indexed biomedical papers, and help us develop algorithms in other areas.

Keywords: Finding semantic relations Indexed papers Coordinating rules Threshold

chinaXiv:201711.02059v1